Университетский менеджмент в Интернет

***О.В.Леонова**, ведущий математик ЦКТ Уральского государственного университета.*

ПОИСК В ИНТЕРНЕТ. РУССКОЯЗЫЧНЫЕ СИСТЕМЫ ПОИСКА

Сегодня в России уже несколько десятков тысяч серверов, а число пользователей, работающих в режиме on-line, превысило стотысячный рубеж.

Все эти компьютеры предоставляют информационный сервис, который компании или отдельные граждане могут использовать в работе и повседневной жизни, например, поисковые системы и базы данных или электронные формы для заказа товаров. Основной вопрос, который сегодня стоит перед пользователями Интернет, — как найти и получить необходимую информацию.

За годы развития Интернет были разработаны различные средства доступа к информации. Это такие виды сетевого взаимодействия как

- FTP, Gopher — системы передачи информации
- Archie, WAIS, Veronica — системы поиска информации в сети
- Telnet, E-mail, UseNet, IRC — коммуникационные сервисы
- WWW (World Wide Web) — мультимедиа система

World Wide Web («Всемирная паутина») возникла в 1994 году в CERN (Европейская лаборатория физики элементарных частиц). Ее появление было вызвано необходимостью единообразного способа доступа к различным видам информации (текстам, графическим изображениям, звуковым фрагментам), не требуя при этом выполнения большого количества действий и специальной подготовки со стороны пользователя.

Для работы с системой WWW вам необходимо установить на своем компьютере WWW-browser (WWW-браузер) — специальную программу просмотра. **Браузер** — это программа — клиент, которая взаимодействует с системой WWW, получает затребованные вами документы и отображает их на экране. Документы, используемые системой WWW, называются гипертекстовыми документами. **Гипертекст** — это текст, который внутри себя содержит **ссылки** на другие документы. При подготовке документов для WWW используется специальный язык HTML (HyperText Markup Language — язык разметки гипертекста). HTML — стандарт, который представляет собой набор команд, описывающих структуру документа. Конкретный вид документа определяет программа-браузер, которая интерпретирует HTML-документ и отображает его на экране в отформатированном виде. Команды HTML вставляются в текст и определяют, наряду с внешним видом документа, логический статус отдельных фрагментов текста. Например, среди команд HTML есть команда для выделения названия документа (<TITLE>), есть команды для выделения заголовков различных уровней внутри документа (<H1>, <H2>, <H3>, ...), есть команды, позволяющие вставить в документ другие объекты (изображения, звуки, анимацию), команды, с помощью которых устанавливаются гипертекстовые связи с другими документами — ссылки (<A>) и т.д.

С помощью WWW-браузера возможно пользоваться также другими сервисами Интернет. Например, два наиболее популярных сейчас браузера Netscape Navigator и Microsoft Internet Explorer позволяют обращаться к анонимным Gopher-, Wais-, FTP-серверам,

серверам телеконференций UseNet, пользоваться электронной почтой (E-mail), а также дают возможность доступа к удаленному компьютеру по протоколу Telnet.

По мере развития Интернет увеличивается объем информации в ней содержащейся и вместе с этим возникает проблема поиска нужной информации. Таким образом, вероятность существования необходимой информации возрастает, а возможность ее нахождения уменьшается. Теоретически гипертекстовая природа WWW обеспечивает нахождение любой информации в процессе целенаправленного продвижения по ссылкам. Однако, среди более 60 млн. документов (а именно столько документов, по некоторым оценкам, существует сегодня в Интернет), найти нужный документ, продвигаясь от ссылки к ссылке, практически невозможно.

Перед тем как перейти к вопросу о том, **как** правильно искать в Интернет нужный вам документ, необходимо разобраться в том, **где** искать. Прежде всего, необходимо классифицировать информационные ресурсы Интернет. ***По способу представления информации*** все информационные ресурсы можно разделить следующим образом:

- Web-ресурсы
- Базы данных
- Файловые серверы
- Телеконференции (UseNet)
- Gopher-серверы

Все чаще WWW интерфейс используется как стандартный метод доступа к остальным ресурсам. Методы поиска информации могут быть различны. Как уже отмечалось, есть возможность искать необходимую вам информацию переходя от ссылки к ссылке, т.е. **вручную**. Однако, учитывая размеры Интернет, можно предположить, что вероятность найти нужный документ очень низкая. Лучший вариант — воспользоваться специально предназначенным для этого **сервером** Интернет. Сервер — это компьютер, программа, а также набор данных. Сервер (или сайт) обеспечивает определенный сервис в Интернет. Здесь можно провести аналогию с поиском книги в библиотеке. Для того, чтобы книгу или статью легко было найти, ей присваивается уникальный идентификатор, состоящий из букв и цифр. Таким образом, зная название книги, библиотекарь легко найдет ее среди бесчисленного множества других. Поисковый сервер занимается тем, что собирает данные в Интернет, а затем позволяет этими данными воспользоваться. Сегодня поисковых серверов насчитывается свыше 120. Наиболее полный их список есть по адресу http://ugweb.cs.ualberta.ca/~mentor02/search/search-all.html. Остается только выбрать, какому из них отдать предпочтение.

Чтобы определить, на каком поисковом сервере остановить свой выбор, необходимо знать, как организован сбор информации для этих серверов. Для того, чтобы поисковая система отвечала своему назначению, информация должна быть предварительно накоплена и просмотрена. Есть два основных *способа сбора информации* для систем поиска и связанных с ними *способа организации* собранной информации.

1. Первый способ — **ручной сбор информации** — означает, что все документы последовательно просматриваются группой специалистов. Такой подход предполагает организацию поисковой системы как **предметно-ориентированной**, где информация по определенным темам собрана в соответствующих каталогах. Примерами таких каталогов являются: ***Yahoo!*** (http://www.yahoo.com/), ***Magellan*** (http://www.mckinley.com/) — среди зарубежных каталогов; ***Созвездие Интернет*** (http://www.stars.ru/), ***Russia on the Net*** (http://www.ru/), ***«Ау!»*** (http://www.rocit.ru/) — среди российских каталогов.

Этот подход требует очень большой доли труда квалифицированных специалистов. Однако документы, просмотренные и разобранные таким образом, более адекватны теме.

2. **Сбор информации с помощью роботов (search robots)**. В этом случае поисковая система представляет собой **Search Engine (SE) — машину поиска**. Вся предварительная работа по просмотру документов выполняется поисковым роботом. Робот — это программа, которая автоматически просматривает структуру всех гипертекстовых ссылок и *индексирует* содержимое всех обнаруженных по ссылкам документов. При индексации фиксируются положения всех более или менее значащих слов, которые называются *ключевыми* (к «неключевым» словам относятся союзы, предлоги, местоимения и т.д.). После разбора документа робот включает его в свою базу данных. В данном случае пользователь будет иметь дело с SE, обращаться к базе данных которой можно только посредством специального интерфейса.

Информация, собранная роботом, имеет больший объем, чем при ручном сборе, поскольку количество документов, которые

просматривает робот, может быть любым. Однако в этом случае формальным критерием оценки документов служат отдельные слова, а также то, как часто они встречаются в документе, в какой части документа они находятся и т.д. в зависимости от алгоритма, а не общий смысл документа. Поэтому, разные по смыслу документы могут быть объединены по формальным признакам. По этой причине среди найденных документов может быть много совершенно не относящихся к теме поиска. В этом отличие SE от ручного сбора информации.

Общее количество известных программ-роботов уже превышает 150. Каждый робот использует свой алгоритм просмотра и индексации документов, поэтому информация, накопленная двумя разными роботами, может быть различна. Это означает, что использование одних и тех же ключевых слов в различных SE приведет к разным результатам. Важно знать также, что с помощью SE возможен поиск как среди HTML-документов на WWW-серверах, так и среди других типов документов и на других типах серверов.

Рассмотрим самые популярные машины поиска. Качество, а значит, и популярность поисковой машины определяются несколькими параметрами:

- размером базы данных SE (т.е. пространством проиндексированных документов)
- процедурой создания запросов к данной SE
- характером выдаваемой информации (ранжирование, фрагменты текста, краткое содержание и т.п.)
- скоростью обработки запроса
- обратной связью (возможность уточнения результатов поиска)

Российские системы поиска

Для поиска документа на русском языке лучше воспользоваться русской поисковой системой. Если известна тематика искомого документа или можно оценить, на каком сервере он может находиться, но неизвестен адрес этого сервера, тогда лучше будет использовать какой-нибудь тематический каталог (или, рубрикатор). Пользоваться таким предметным каталогом несложно. Рассмотрим один из каталогов в русской части Интернет.

«СОЗВЕЗДИЕ ИНТЕРНЕТ» (http://www.stars.ru/)

Слева на экране находятся темы, по которым рассортированы все ресурсы, зарегистрированные в каталоге: поисковые сервисы и каталоги, компьютеры и технологии, экономика и бизнес, политика и право, культура и искусство, образование и наука, средства массовой информации, техника и транспорт, медицина и здоровье, отдых и развлечения, разное. Кроме того, вы можете воспользоваться быстрым поиском. Дело в том, что при регистрации ресурса в этом каталоге для каждого ресурса вводятся слова с его описанием. На первой же странице каталога появляется строка для ввода с предложением ввести **слово в описании ресурса**. Наберите, например, в поле для ввода слово *управление*, и нажмите кнопку ПОИСК. На экране появится сообщение о том, сколько ресурсов, содержащих в описании слово *управление*, имеется в данном каталоге. Ниже будет выведена таблица (по 10 ресурсов на страницу), в левой части которой название сервера и ссылки на первую страницу данного сервера, или несколько первых страниц для различных кодировок: Win (windows-1251), KOI (koi8-r) — кодировки русского языка, Eng (english) — английского языка.

Теперь рассмотрим, как пользоваться поисковой машиной (SE). Для того чтобы эффективно пользоваться SE, необходимо помнить, что на каждой поисковой машине существует свой язык запросов к накопленной ею базе данных. Поэтому, зайдя на поисковый сервер, прежде чем формировать запрос, надо посмотреть ссылку «Помощь» (или «Help») с описанием порядка формирования запросов. В этой статье приводится описание языка запросов для нескольких поисковых машин. Отличием русских поисковых машин является то, что с их помощью, в отличие от иностранных SE, можно искать документы, набирая русские ключевые слова в поле запроса. Особенности морфологии русского языка накладывают определенные требования на SE, которые используются для индексации русской части Интернет. Если в английском языке достаточно поменять окончание, чтобы найти различные варианты одного и того же слова, то в русском языке может изменяться все слово целиком. С этим связаны трудности индексации и поиска русских документов в Интернет.

МАШИНА ПОИСКА RAMBLER (http://www.rambler.ru)

Данная система служит для поиска документов на серверах России и стран СНГ. В ее

базе данных содержится более 2,000,000 документов (адресов URL[1]) с более чем 15,000 хостов (имен DNS[2]). Имеет развитый язык запросов и гибкую форму вывода результатов. Однако морфологический разбор слова не производится.

ПРОСТОЙ ЗАПРОС

В простом запросе вы можете использовать одно или несколько слов, разделенных пробелами. Могут быть использованы как русские, так и английские словосочетания. По умолчанию, если вы не используете расширенный поиск и не отметили в нем, что должно встретиться любое слово, считается, что в найденных документах должны содержаться все слова. После того, как вы ввели ключевые слова, нажмите правой кнопкой мыши на надписи «Поиск», которая расположена справа. Кроме простого ввода слов, вы можете использовать язык запросов, принятый для поиска документов на «Rambler». В этой таблице коротко описаны элементы этого языка.

Элементы	Пояснение	Примеры
Логические связки: And, Or, Not.	Поисковые термины могут быть объеденены логическими операциями посредством служебных слов And, Or и Not. Символы '&', '|' и '!' могут использоваться в сочетании со служебными словами или вместо них.	*Управление and законодательство not бюджет* Во всех найденных документах будут присутствовать слова *управление и законодательство* и отсутствовать слово *бюджет.*
Регистр букв	Любой поисковый термин может содержать в себе как заглавные, так и прописные символы. Индекс базы данных строится с приведением слов к прописным символам.	*Федеральный бюджет* или *федеральный Бюджет* Будут найдены одни и те же документы.
Усечение слов * и ?	Возможно использование метасимволов '*' и '?' для обозначения произвольной части слова и произвольного символа слова. По умолчанию система ищет документы с теми ключевыми словами, которые вы ввели.	*орган?зация* and *управлен*and *ВУЗ* Знак ? используется, если нет уверенности в написании слова. Знак * заменяет несколько букв слова.
Весовые коэффициенты + и -	Вы можете использовать '+' и '-' для увеличения/уменьшения весового значения любого слова. Возможно многократное использование данных символов.	*-система* and *++управлен** Слово *система* будет иметь меньший вес, поэтому документы с этим словом будут расположены после слов, начинающихся на *управлен*
Поиск в части документа $ спец. слово	Для этого вы можете использовать специальные слова: $All (используется по умолчанию), $URL, $Title, $Header, $Essence, $Address. Специальные слова начинаются с символа '$'.	*$TITLE: управление and $URL: virlib.eunnet.net* Будут найдены документы, у которых в поле заголовка есть слово *управление* и они содержат ссылку на сервер с адресом http://virlib.eunnet.net/
Логические группы ( )	Термины могут быть сгруппированы посредством использования символов '('и')'. Возможна многократная вложенность скобок в сочетании с логическими операторами.	*управленческие* and *(функции* or *полномочия)*

Вывод результатов поиска.

На одну страницу будет выведено15 первых из всех найденных документов, а внизу страницы (если общее число найденных документов больше 15) появится строка со ссылками на страницы с остальными найденными документами: по 15 документов на страницу. «Rambler»производит ранжирование найденных документов в зависимости от частоты употребления и местоположения искомых слов. В начале списка будут выведены документы, наиболее полно удовлетворяющие запросу. После заголовка документа, который одновременно является ссылкой на данный документ, в скобках будет стоять число — 1,0000, что означает максимальное соответствие запросу, и ниже. Далее следует несколько первых строк документа, его адрес в явной форме, дата его создания или модификации, объем файла, в скобках вид кодировки. Если адресов у документа несколько, это означает что, либо найдены полностью идентичные документы, либо это один и тот же документ, но в разных кодировках.

ДЕТАЛЬНЫЙ ЗАПРОС

Механизм составления детального запроса реализован через меню.

Ключевые слова набираются в поле запроса через пробел. Под строкой для ввода ключевых слов можно выбрать позиции для поиска.

* **Поиск в:** Российский Web, Российский Usenet, имена URL (адреса), название документов, заголовках документов, начале документов, поле адресов. Выбрав одно из полей, можно ограничить область поиска документа: www серверами; телеконференциями Usenet; адресами серверов Интернет; именами файлов; полями <TITLE> в гипертекстовых документах; первыми абзацами документов.

* **Кол-во**: 15, 30, 50. Количество результатов, которые будут выводиться на одну страницу.

* **Слова.** Логические операции над ключевыми словами. Опция «**Все**» означает, что в каждом найденном документе будут все ключевые слова (аналог and и &). «**Любое**» означает, что в каждом найденном документе будет присутствовать хотя бы одно из ключевых слов (аналог or и |).

* **Форма вывода** результатов. Нормальная форма (используется по умолчанию при простом запросе): заголовок, показатель соответствия запросу (числовой и в виде точек), первые строки документа, URL документа, дата создания, объем, кодировка. Краткая форма: заголовок, степень соответствия запросу. Детальная форма: более подробная информация о документе, например, перечислены все заголовки, а также когда документ последний раз проверялся роботом.

* **Расширить слова.** Опция «**нет**» означает, что искать надо строго по введенным ключевым словам, не добавляя окончаний. «**Да**» — добавить к введенным ключевым словам все возможные окончания (аналог *).

* **От даты**: **До даты**: Например, От даты: 21/Mar/96 До даты: 1/Jan/98. Будут найдены документы, созданные или модифицированные в период с 21 марта 1996 г. до 1 января 1998 г.

* **Исключить документы, содержащие следующие слова.** Слова, которые будут введены в этом поле, будут отсутствовать в найденных документах.

* **Сайт или часть URL, в которых произвести поиск.** Можно ограничить поиск только одним сервером (сайтом), набрав в этом поле его URL или несколькими сайтами, введя только часть URL, а не искать во всей базе данных поисковой машины. Например, www.stack.net, gopher://gopher.dux.ru/, ua.

Главный недостаток «Rambler» — невозможность осуществлять поиск по целой фразе или хотя бы указывать в запросах предельное расстояние между искомыми терминами. Случайное сочетание совершенно не связанных слов, например, в начале и конце текста, приводит к выдаче ссылок на документы, совершенно не релевантные запросу. Несовершенный метод ранжирования результатов по степени соответствия запросу приводит к тому, что искомые документы часто оказываются не в начале списка.

«АПОРТ!» (http://www.aport.ru/)

Поиск ведется по 1 327 132 документам (2 759 935 URL, 10 971 сервер). Это данные на 1998-02-28. Вы можете набрать интересующие вас ключевые слова через пробел. Машина найдет все документы, в каждом из которых содержатся все введенные слова. Важное достоинство «Апорт» — поиск с учетом морфологии русского языка. Вы можете вводить слова в любой грамматической форме. Например, запрос *университетское управление* будет полностью эквивалентен запросу *университетским*

управлением. Кроме того, английские слова могут указываться в запросе наравне с русскими.

В таблице — краткое описание языка запросов поисковой машины «Апорт».

слов, вы можете выбрать эту опцию. Машина автоматически исправит ошибки.

× **Очистить историю запросов**. Все предыдущие запросы сохраняются.

Логические операторы: ***и, или***	Оператор *и* подразумевается (т.е. действует по умолчанию), его можно опускать: запрос *университетское управление* полностью эквивалентен *университетское и управление*. По любому из этих запросов будут найдены документы, содержащие оба слова. По запросу *университетское или управление* будут найдены документы, содержащие хотя бы одно из указанных слов.
Двойные кавычки *« »*	Двойные кавычки следует использовать, если вы хотите искать словосочетание. По запросу *«университетское управление»* будут выданы только документы, содержащие указанное словосочетание (возможно, в разных грамматических формах), тогда как по запросу *университетское управление* будут выданы и те документы, где заданные слова стоят далеко друг от друга и, может быть, даже в обратном порядке.
Круглые скобки *( )*	Круглые скобки задают порядок действия логических операторов. По запросу *быстрый или качественный поиск* будут выданы документы, содержащие либо слово «*быстрый*», либо одновременно слова «*качественный*» и «*поиск*» (оператор и действует первым). По запросу *(быстрый или качественный) поиск* будут выданы документы, где встречаются одновременно слова «*быстрый*» и «*поиск*», либо «*качественный*» и «*поиск*».
Фигурные скобки { }	Фигурные скобки ограничивают расстояние между словами, задавая его числом предложений. Запросу {3, *управленческие функции*} будут соответствовать документы, где слова « *управленческие*» и «*функции*» встречаются в пределах трех соседних предложений. Цифра (вместе с запятой) может опускаться, тогда подразумевается 1, то есть слова должны встречаться в одном предложении: {*управленческие функции*}.
Квадратные скобки [ ]	Квадратные скобки аналогичны фигурным с той лишь разницей, что расстояние между словами измеряется не в предложениях, а в словах. По запросу [4, уголовные преступления] будут найдены документы, где между словами стоит не более двух посторонних слов.

Для поиска по URL используйте оператор URL (в форме URL: или URL=). Если надо найти упоминания адреса сервера в текстах документов, рекомендуется использовать поиск в пределах предложения с заменой '/' на пробелы.

Например, {*UniMgmt.EUNnet.net unimng*}.

Не используйте в запросе так называемые «стоп-слова». К «стоп-словам» относятся предлоги, союзы, междометия и т.д. Если вы укажете в запросе слово *пожалуйста*, то «Апорт» не найдет никаких документов.

Дополнительные возможности.

- **Исправлять ошибки в запросе**. Если вы не уверены в правильности написания ключевых

× **Форма результата**. Предлагается возможность гибкого указания формы выдачи результатов поиска.

× **Перевод запроса.** Автоматического перевода запроса с русского на английский и наоборот. В поисковую строку можно ввести термины на любом из двух языков и выбрать из меню условие: искать только на английском, на английском и русском, только на русском.

× **Перевод результата**. Возможно указать необходимость перевода результатов на английский, русский, либо не переводить.

Результат поиска.

По 10 на страницу. Название документа,

дата создания, ссылка на документ в явном виде (URL документа), кодировка, степень соответствия запроса (в процентах), количество предложений, соответствующих запросу. Есть возможность посмотреть на реконструкцию текста (т.е. не весь текст, а только его реконструкция). «Апорт!» показывает фрагмент текста, который удовлетворяет искомому запросу.

Недостатком «Апорт!» является невозможность управлять ранжированием результатов.

ЯNDEX
(http://yandex.ru/)

Проанализировано 12043 серверов. Накоплена информация о 2 402 168 ссылок (URL). Область поиска этой SE — «русская Интернет», т.е. домены верхнего уровня 'su' и 'ru', домены бывшего СССР (например, 'ua', 'kz') и Web-сайты в других доменах, содержащие русские тексты. «Яndex» «понимает» русскую морфологию и различные русские кодовые таблицы. Учитывает при разборе ключевых слов морфологию русского языка. В русском языке возможно изменение слова в целом, а не только его окончание.

ПРОСТОЙ ПОИКС.

При заходе на сервер этой SE в окне браузера появляется окошко для ввода запроса.

Естественный язык.

Поскольку использование специального языка запросов требует некоторого навыка работы с SE, очень важно, что «Яndex» предоставляет возможность свободного запроса, то есть вы можете набрать запрос на естественном языке. В этом случае вы тоже получите документы в той или иной степени удовлетворяющие запросу.

Специальный язык запросов.

В том случае, если удовлетворяющие вас документы не найдены по запросу на естественном языке, вы можете воспользоваться специальными символами для формирования запроса. Внизу поля для ввода запроса имеется надпись: «строгий поиск (с языком запросов)». Если вы поставите флажок напротив этой надписи, то все символы этого языка запросов могут быть использованы.

Независимо от того, в какой форме вы употребили слово в запросе, поиск учитывает все его формы по правилам русского языка. Например, если задан запрос *идти*, то в результате поиска будут найдены ссылки на документы, содержащие слова *идти*, *идет*, *шел*, *шла* и т.д. На запрос *окно* будет выдана информация, содержащая и слово *окон*, а на запрос *отзывали* — документы, содержащие слово *отозвали*.

Кроме того, возможен поиск с указанием желаемого расстояния между словами. Если все слова в тексте перенумеровать по порядку их следования, то расстояние между словами a и b — это разница между номерами слов a и b. Таким образом, расстояние между соседними словами равно 1 (а не 0), а расстояние между соседними словами, стоящими «не в том порядке», равно -1. То же самое относится и к абзацам. В таблице приведен язык запросов к поисковой машине «Яndex».

Элементы	Пояснение	Примеры
Заглавные буквы	Если в запросе набрано слово с большой буквы, будут найдены только слова с большой буквы, в противном случае будут найдены как слова с большой, так и с маленькой буквы.	Например, запрос *вуз* (также как и *ВУЗ*) найдет любое упоминание этого слова. Запрос *Вуз* — только те случаи, когда слово написано с большой буквы.
Точная словоформа '!'	По умолчанию поиск учитывает все формы заданного слова согласно правилам русского языка. Однако существует возможность поиска по точной словоформе, для этого перед словом надо поставить восклицательный знак '!'.	Так по запросу '*! управленческих*' будут найдены все документы, содержащие словоформу *управленческих*, а по запросу '*управленческие ~~ ! управленческих*' — документы, в которых есть слово *управленческие*, кроме тех, которые были найдены по первому запросу.
Логическое сложение &	Несколько набранных в запросе слов, разделенных пробелами,	Например, при запросе '*документооборот управление*' (или

	означают, что каждое из них должно входить **в один абзац** искомого документа. Тот же самый эффект произведет употребление символа '&'.	'*документооборот & управление*'), результатом поиска будет список документов, в которых в одном абзаце содержатся и слово '*документооборот*', и слово '*управление*'.
&&	Двойной оператор && ищет также как и &, но **во всем документе**.	По запросу '*документооборот && управление*' будут найдены документы, содержащие где бы то ни было оба эти слова
Логическое вычитание | или ,	Между словами можно поставить знак '|' (или запятую ','), чтобы найти документы, содержащие любое из этих слов.	Запрос вида '*функции | полномочия*' или '*функции, полномочия*' задает поиск документов, содержащих в одном абзаце хотя бы одно из слов *функции* или *полномочия*.
Логическое отрицание ~	Этот знак, тильда ~, позволит найти документы с **абзацем**, содержащим первое слово, но не содержащим второе.	По запросу '*централизация ~ децентрализация*' будут найдены все документы, содержащие слово '*централизация*', рядом с которым (в пределах абзаца) нет слова '*децентрализация*'.
~~	Двойной оператор ~~ ищет **в пределах документа.**	Запрос '*централизация ~~ децентрализация*' выдаст все документы со словом '*централизация*', но без слова '*децентрализация*'
/n	Если между двумя словами поставлен знак '/', за которым требуется, чтобы расстояние между ними не превышало этого числа слов.	Например, задав фразу '*систе* сразу напечатано число, значит, *ма /2 управления*', Вы требуете найти документы, в которых содержатся и слово 'холодный' и слово 'вода', причем расстояние между ними должно быть не более двух слов и они должны находиться в одном абзаце.
/+n	Если порядок слов и расстояние точно известны, можно воспользоваться пунктуацией /+n. Так, например, задается поиск слов, стоящих подряд.	Запрос '*система /+1 управления*' означает, что слово 'вода' должно следовать непосредственно за словом 'холодный'. (Кстати, к тому же результату приведет запрос 'холодная вода')
Ограничение по расстоянию /(n m)	В общем виде ограничение по расстоянию задается при помощи пунктуации вида '/(n m)', где 'n' минимальное, а 'm' максимально допустимое расстояние. Отсюда следует, что запись '/n' эквивалентна '/(-n +n)', а запись '/+n' эквивалентна '/(+n +n)'.	Запрос '*система /(-2 4) управления*' означает, что '*управления*' должна находиться от '*система*' в интервале расстояний от 2 слов слева до 4 слов справа.

	Практически все знаки можно комбинировать с ограничением расстояния.	Например, результатом поиска по запросу *система ~ /+1 управления* будут документы, содержащие слово '*система*', причем в этих документах слово '*управления*' не следует непосредственно за словом '*система*'.
	Когда знаки ограничения по расстоянию стоят после двойных операторов, то употребленные там числа — это расстояние не в словах, а в абзацах. Расстояние в абзацах определяется аналогично расстоянию в словах.	Запрос *'система && /1 управления'* означает, что слово 'вода' должно находиться в том же самом, либо в соседнем со словом 'холодный' абзаце.
Круглые скобки ( )	Вместо одного слова в запросе можно подставить целое выражение. Для этого его надо взять в скобки.	Например, запрос '(*организация, система*) /+1 (*управления|менеджмента*)' задает поиск документов, которые содержат любую из фраз '*организация управления*', '*организация менеджмента*', '*система управления*', '*система менеджмента*'.
$Title:	Можно искать информацию в заголовках (имя «зоны»: Title) и ссылках (имя «зоны»: A). Синтаксис: $имя_зоны логический_множитель	Запрос '$Title КомпТек' ищет в заголовках документов слово 'КомпТек'.
$A:	Можно искать информацию в ссылках.	
$A логическое выражение или $Title логическое выражение	Можно использовать логические операторы после $A или $Title	Запрос '$A (КомпТек | Dialogic)' находит документы, в ссылках внутри которых есть одно из слов 'КомпТек' или 'Dialogic'.

Ранжирование результатов поиска

При поиске для каждого найденного документа «Яндекс» вычисляет величину релевантности (соответствия) содержания этого документа поисковому запросу. Список найденных документов перед выдачей пользователю сортируется по этой величине в порядке убывания. Релевантность документа зависит от ряда факторов, в том числе от частотных характеристик искомых слов, веса слова или выражения, близости искомых слов в тексте документа друг к другу и т.д.

Пользователь может повлиять на порядок сортировки, используя операторы веса и уточнения запроса. **Задание веса слова или выражения** применяется для того, чтобы увеличить релевантность документов, содержащих «взвешенное» выражение.

Синтаксис: слово: число

или (поисковое_выражение):число

Например, по запросу 'поисковые механизмы:5' будут найдены те же документы, что и по запросу 'поисковые механизмы'. Разница состоит в том, что наверху списка найденного окажутся документы, где чаще встречается именно слово 'механизмы'. Запрос 'поисковые (механизмы|машины|аппараты):5' равнозначен запросу 'поисковые (механизмы:5|машины:5|аппараты:5)'.

Задание уточняющего слова или выражения применяется для увеличения релеватности документов, содержащих уточняющее выражение.

Синтаксис: <- слово

или <- (уточняющее_выражение)

Например, по запросу 'компьютер <- телефон' будут найдены все документы, содержащие

слово 'компьютер', при этом первыми будут выданы документы, содержащие слово 'телефон'. Если ни в одном документе со словом 'компьютер' нет слова 'телефон', результат запроса будет эквивалентен запросу 'компьютер'.

Результаты поиска.

Результаты поиска появляются на экране по 10 на страницу по мере убывания степени соответствия запросу (максимальная степень соответствия — [1.000000]). Внизу каждой страницы находятся ссылки (по номерам) на другие страницы с найденными по запросу документами. Для каждого документа в списке найденного указан его заголовок, ссылающийся на **размеченный документ**, начало текста документа, кодировка, размер в байтах, дата и URL документа, ссылающийся на оригинальный документ. Если вы не хотите, чтобы результаты запроса пропадали с экрана, вы можете нажать на маленькие окошечки слева от явной ссылки на оригинальный документ. При этом документ загружается в новое окно браузера. При нажатии на явную ссылку оригинальный документ загрузится в текущее окно браузера.

Что означает разметка документа? Если в списке найденного нажать на заголовок документа, Вы увидите так называемую «подсветку». «Яndex» при индексации запоминает положение слова в документе, что дает возможность выделить (подсветить) слова, найденные в тексте. И не просто подсветить, а переходить с одного слова на другое. При этом подсвечиваются не все слова, входящие в запрос, а только те, которые удовлетворяют поисковому выражению.

Слова выделены угловыми стрелочками. Каждая стрелочка ссылается на следующее или предыдущее «найденное» слово. Чтобы увидеть первое найденное слово, нажмите на стрелочку *влево* , чтобы увидеть последнее — на стрелочку *вправо*. Переход на следующее слово — стрелочка > справа от слова, переход на предыдущее — слева <. Первое и последнее слова указывают на верхнюю и нижнюю таблицу соответственно. В начале размеченного документа помещается табличка с ссылками на первое и последнее найденное слово и на оригинальный документ. В конце документа — аналогичная табличка, где приводится статистика, то есть — сколько слов найдено (подсвечено) в данном документе. Если файлы были изменены, а индекс по ним не обновлен, об этом выдается соответствующее предупреждение.

Можно **ограничить область поиска**, отметив «искать в найденном» на странице результата.

Если же удовлетворяющий вас документ не найден, есть еще возможность воспользоваться **поиском документов по образцу**. Для этого нажмите на надпись «Найти похожие документы», которая находится под наиболее удовлетворяющим вас документом. При этом будет сформирован новый запрос к поисковой машине «Яndex» и найденные документы будут походить на исходный. Однако этой опцией надо пользоваться аккуратно, поскольку количество документов, найденных в результате может превысить разумный предел и, следовательно, не приведет ни к чему.

АКАДЕМИЧЕСКИЙ ПОИСК.

Нажав левой клавишей мыши на надпись «Advanced», расположенную в правой части экрана вместе с другими пунктами меню, на экране вы получите поле для ввода запроса и меню:

× **Уточнение** запроса. Если вы введете слова в этом поле, то первыми документами в списке результатов будут документы, содержащие эти слова.

× **Выдача** результатов. Здесь можно выбрать краткую (заголовок и степень соответствия запросу) либо стандартную (которая была описана выше) форму выдачи результатов, а также количество документов, выводимых на страницу (10, 20 или 50).

× **Зона поиска.** Искать во всем документе, только в заголовках, только в ссылках.

В остальном этот раздел ничем не отличается от простого поиска с «Яndex», т.е. в поле запроса можно использовать как естественный язык, так и специальный язык запросов, пометив пункт «строгий поиск (с языком запросов)».

Кроме прямого использования «Яndex», есть возможность сформировать с ее помощью запрос и отправить его на поисковые машины «AltaVista» или «Rambler». Для каждой из этих SE у «Яndex» есть специальный интерфейс, где пользователь набирает ключевые слова, отмечает необходимые для поиска опции. Нажав на кнопку «Обработка запроса», вы передаете свой запрос на «Яndex», которая обрабатывает его с учетом морфологии русского языка и отправляет на «AltaVista» или «Rambler» (в зависимости от выбранного вами интерфейса). Интерфейсы написаны для двух кодировок русского языка: Windows-1251 или KOI8-R

Интерфейс «Яndex» для «Rambler» (http://www.comptek.ru/ramb.html)

• **Учет словосочетаний.** Если поле не помечено каждое слово заменяется на все свои формы, т.е. реализуется ***морфологический режим***

обработки запроса. Если поле помечено, по возможности учитываются синтаксические связи между словами в запросе, т.е. реализуется ***морфосинтаксический режим обработки запроса.***

• **Режим**. Режим «Поиск» — запрос посылается на «Rambler». Если выбран режим «разбор запроса», то при нажатии на кнопку «ПОИСК!» на экран выдаётся протокол морфологического анализа всех слов запроса (из поля «Запрос»). Для каждого слова приводятся все варианты его морфологического разбора. Для каждого варианта разбора указаны все его грамматические характеристики. Если слово отсутствует в словарях системы, то она генерирует гипотетическую модель словоизменения этого слова. В конце протокола приводится расширенный запрос, сгенерированный словарным сервером.

• **Поиск в** WWW, UseNet, именах URLs (указывает на область поиска)

• **Операции со словами. Все** — означает логическую операцию И. **Или** — логическая операция ИЛИ.

• **Количество результатов** на страницу (10, 20 и т.д.)

• **Форма вывода** (нормальная, краткая, детальная)

Следующие поля не являются обязательными и применяются только для поиска в WWW (использование этих полей может замедлить поиск). (Вы не можете использовать метасимволы '*' и '?' в следующих полях)

•**От даты: До даты: формат 21/Mar/96**. Дата последнего изменения искомых документов.

•**Исключить документы, содержащие следующие слова**.

•**Сайт или часть URLs, в которых произвести поиск**. Примеры: 'www.stack.net' 'gopher://gopher.dux.ru/' 'ua'

Запрос задается в формате детального запроса Rambler.

Морфологический режим обработки запроса.

В этом режиме каждое слово из запроса заменяется на все свои формы — с учётом родов, чисел, склонений, спряжений. Учитывается также омонимия (напр. по слову «раздел» будут даны все формы глагола «раздевать» и существительного «раздел»). Если Вы хотите искать слово только в той форме, в которой Вы его задали, поставьте его в кавычки. Слова, заключённые в квадратные скобки, трактуются как словосочетание, то есть часть запроса (их может быть несколько), взятая в квадратные скобки, обрабатывается в морфосинтаксическом режиме (как запрос при помеченном поле «Учет словосочетаний»). Вложенность квадратных скобок не допускается.

Морфосинтаксический режим обработки запроса.

Реализуется при помеченном поле «Учет словосочетаний» для всего запроса, или для частей запроса, взятых в квадратные скобки, когда это поле не отмечено. В этом режиме поисковый запрос трактуется как фраза на естественном языке. При этом поиск становится более релевантным, поскольку находится гораздо меньше «мусора», так как учитываются синтаксические связи между словами запроса. Также происходит частичное снятие омонимии: например, в случае задания поисковой фразы *после проверки* предлог *после* не будет считаться формой слова *посол* и последнее не будет дано для поиска во всех формах.

Запрос обрабатывается следующим образом:

◆ Все слова из запроса должны находиться в искомых документах, поэтому при генерации расширенного запроса применяется оператор and (&).

◆ Если слова в запросе синтаксически связаны, то расширенный запрос строится с учетом синтаксических связей.

◆ Знаки препинания игнорируются.

◆ Слова, набранные латиницей, в том числе and,or,near,not, считаются составной частью фразы (а не операторами языка запроса).

В данный момент учитываются два вида синтаксической связи:

1. Согласование существительного с прилагательным или причастием в роде, числе и падеже.

Например, если задан запрос

информационные технологии, то расширенный запрос будет выглядеть следующим образом: ((информационная & технология) | (информационной & технологии) | (информационную & технологию) | ((информационной | информационною) & (технологией | технологиею)) | (информационные & технологии) | (информационных & (технологий | технологиях)) | (информационным & технологиям) | (информационными & технологиями)), т.е. существительное и прилагательное согласованы в роде, числе и падеже. «Морфологическое» расширение этого запроса выглядело бы так: (информационная or информационной or... /*далее по всем падежам и числам*/) & (технология or технологии or... /*по всем падежам и числам*/).

2. Управление предлога существительным или именной группой. Например, запрос *документооборот в управлении* приводит к генерации расширенного *запроса (документооборот or документооборота or...* /*по всем падежам и

числам*/) near *в* near (*управлении* or *управлениях*)

Интерфейс «Яndex» для «AltaVista» (http://www.comptek.ru/alta.html)

Используя этот интерфейс, «Яndex» посылает ваш запрос на поисковую систему «AltaVista», предварительно его обработав. «AltaVista» имеет русский интерфейс, но поиск с помощью этого интерфейса не учитывает морфологии русского языка. Однако эта SE обладает огромной базой данных проиндексированных документов, поэтому использование «Яndex» для формирования запроса в сочетании с большим количеством документов может дать хороший результат.

• **Учет словосочетаний** аналогичен такому же пункту для «Rambler».

• **Область поиска:** WWW, UseNet, Россия (домены 'su' и 'ru'), Россия и США (домены 'su', 'ru', 'com', 'edu', 'org').

• **Вывод результата.** Стандартная, компактная, детализация, счетчик (будет указано только количество релевантных документов).

Запрос:

• **Наиболее значимые слова.** Слова в этом поле будут восприняты как дополнительные ключевые, кроме того документы, в которых они встречаются, будут располагаться в начале списка результатов.

• **Нач. дата**: **Конеч. дата**: (напр.: 12/Янв/96)

• **Режим** («поиск» или «разбор запроса») аналогичен такому же пункту для «Rambler».

•**Кодировка** (Windows-1251 или KOI8-R)

В таблице приведен пример использования рассмотренных выше SE для поиска информации об университетском управлении.

Машина поиска	Запрос	Результат
Яndex	университетский (менеджмент,управление)	Найдено 111 уникальных документов
	университетский &/2 (менеджмент,управление)	Найдено 14 уникальных документов
Rambler	университетск* and (менеджмент or управление)	Найдено: 926 [676 уникальных]
	университетский and (менеджмент or управление)	Найдено: 130 [106 уникальных]
	университетское & управление	Найдено: 36 [26 уникальных]
	университетское & управление or университетский & менеджмент	Найдено: 53 [43 уникальных]
Апорт!	университетское (управление или менеджмент)	Найдено 989 документов
	{2,университетское управление} или {2,университетский менеджмент}	Найдено 233 документа
	{1,университетское управление} или {1,университетский менеджмент}	Найдено 192 документа

«ЯNDEX». Искомые документы находятся в начале списка. Кроме того, при большом количестве документов возможно уточнение результатов («искать в найденном»).

«RAMBLER». Необходимо отметить, что наличие * в конце слова позволяет «выловить» даже те документы, в которых окончания этого слова были набраны ошибочно. В начале списка много документов, мало относящихся к предмету поиска. Уточнение поиска невозможно.

АПОРТ! Находит слишком много документов, дальнейшее уточнение поиска не предусмотрено. Однако среди первых документов есть документы, относящиеся к теме поиска.

Конечно, с другими ключевыми словами результаты поиска будут отличаться от результатов приведенных здесь.

1. Uniform Resource Locator — адрес документа в Интернет. Например, *http://www.usu.ru/eb-engl.htm.*

2. Domain Name System (доменная система имен) — устанавливает соответствие между компьютером в Интернет и его именем. Система служит для облегчения запоминания имен компьютеров в Интернет.